Inf Retrieval (2009) 12:69–80
DOI 10.1007/s10791-008-9074-8

Evaluation of query expansion using MeSH in PubMed

Zhiyong Lu · Won Kim · W. John Wilbur

Received: 28 March 2008 / Accepted: 3 October 2008 / Published online: 29 October 2008

Abstract This paper investigates the effectiveness of using MeSH® in PubMed through its automatic query expansion process: Automatic Term Mapping (ATM). We run Boolean searches based on a collection of 55 topics and about 160,000 MEDLINE® citations used in the 2006 and 2007 TREC Genomics Tracks. For each topic, we first automatically construct a query by selecting keywords from the question. Next, each query is expanded by ATM, which assigns different search tags to terms in the query. Three search tags: [MeSH Terms], [Text Words], and [All Fields] are chosen to be studied after expansion because they all make use of the MeSH field of indexed MEDLINE citations. Furthermore, we characterize the two different mechanisms by which the MeSH field is used. Retrieval results using MeSH after expansion are compared to those solely based on the words in MEDLINE title and abstracts. The aggregate retrieval performance is assessed using both F-measure and mean rank precision. Experimental results suggest that query expansion using MeSH in PubMed can generally improve retrieval performance, but the improvement may not affect end PubMed users in realistic situations.

Keywords Query expansion · TREC genomics · PubMed · MeSH

1 Introduction

Query expansion refers to the process of reformulating a query to improve retrieval performance, typically by including additional terminology that is synonymous. PubMed, the search engine for over 17 million biomedical and health citations in MEDLINE, allows its users to manually attach search field tags (also called qualifiers) to search terms in square brackets (i.e., search term [tag]) in order to improve on users' original searches. However, such advanced MEDLINE search features were seldom used as seen in the PubMed log analysis (Herskovic et al. 2007). Alternatively, PubMed employs a process called

Z. Lu (✉) · W. Kim · W. J. Wilbur
National Center for Biotechnology Information (NCBI), National Library of Medicine,
Bethesda, MD 20894, USA
e-mail: luzh@ncbi.nlm.nih.gov

Automatic Term Mapping (ATM) that compares and maps untagged terms from the user query to lists of pre-indexed terms in PubMed's translation tables (in this order): the Medical Subject Headings (MeSH) table (mapping search terms to MeSH concepts), the journals translation table (mapping search terms to journal names), and the author index (mapping search terms to author names). In the context of query expansion, we consider only the translation via MeSH here, which is designed by the National Library of Medicine (NLM) for indexing and searching of the MEDLINE database of journal citations. That is, if a user query includes a term that can be mapped to a MeSH concept, the search will then add the MeSH term to the original query. As a result, the ATM process enables the original query to have an access to the MeSH field of indexed MEDLINE documents. For instance, if a user submits a query *tumor* with no search tags attached, the query will be automatically mapped to the MeSH term `Neoplasms` by ATM. Thus, not only documents having the word *tumor* in the title and abstract will be retrieved, any other documents indexed with the MeSH term `Neoplasms` (as well as more specific terms beneath it in the MeSH hierarchy) will be returned as well. Furthermore, individual words in MeSH concepts are also indexed and searchable. Take the previous query for example, documents indexed with MeSH terms that include the word *tumor* (e.g. `Tumor Virus Infections`) will also be returned.

By default, PubMed applies the Automatic Term Mapping feature to every unqualified user query, thus affecting the search results of millions of user queries each day in spite of the fact that many users may not actually realize its existence. It is designed and implemented by the NLM as a means to improve retrieval performance. Unlike the studies in the past that focused on the accuracy of the mapping process (Gault et al. 2002; Carlin 2004; Smith 2004; Shultz 2006), the main objective of this work is to examine the effectiveness of using the MeSH field of indexed documents for improving retrieval performance.

1.1 Related work

There is a significant body of literature on query expansion as an effective technique in the general information retrieval (IR) community (Salton and Buckley 1997; Mitra et al. 1998). However, in the domain of biomedical text retrieval, results have been mixed. Earlier research has suggested the use of this technique could result in elevated retrieval performance (Srinivasan 1996; Aronson and Rindflesch 1997), while no performance gain was seen in others (Hersh et al. 2000).

More recently, this technique has been widely implemented in various retrieval systems built by a number of participating teams in TREC Genomics tracks since 2003 (Hersh and Bhupatiraju 2003). Like previous investigations, there have been some contradictory reports on the benefits of using query expansion. Results from some TREC Genomics *ad hoc* retrieval task participants (Hersh and Bhupatiraju 2003; Abdou et al. 2005) pointed to the use of query expansion through multiple concepts: MeSH headings, substance name fields in MEDLINE, and species as accounting for improved performance. On the contrary, other participating teams (Hersh et al. 2003, 2004, Guo et al. 2004; Bacchin and Melucci 2005) reported no or detrimental impact when expanding original queries with multiple sources including MeSH concepts.

The main differences between the work presented in this paper and the previous studies are (1) the retrieval system being used during evaluation, and (2) the method used for query expansion. All previous analyses based their results on either the vector space model (Salton and Buckley 1988) or probabilistic models like the Okapi model (Robertson and Walker 1994), both of which rank retrieved documents on the basis of their relevance to the queries. In contrast, PubMed is a Boolean search system and displays retrievals in

reverse chronological order by default. Thus, we implemented our own search system that supports both Boolean and probabilistic text retrievals. As a result, not only can we evaluate the query expansion approach in the same setting as used by PubMed, but we can compare retrieval results with term weighting ranking strategies demonstrated in the previous studies.

The second distinction lies in the methods used for query expansion. The expansion terms in general come from two different sources. First, queries can be expanded using thesauri like the Unified Medical Language System (UMLS) Metathesaurus (Lindberg et al. 1993) via either computational approaches (Srinivasan 1996; Aronson 1996; Guo et al. 2004; Abdou et al. 2005) or manual mappings (Hersh et al. 2000). Second, expansion terms can be extracted using retrieval feedback (Srinivasan 1996; Bacchin and Melucci 2005). In our investigation, query expansion was executed through ATM, a unique mapping process of PubMed.

Finally, to the best of our knowledge, this is the first formal evaluation on the benefits of applying the query expansion technique to a daily operational search system as opposed to retrieval systems mainly designed and tested in research laboratories. Therefore, our analysis plays a critical role in the understanding of future technology development needs for PubMed, along with its mission to better fulfill the information needs of millions of PubMed users.

2 Methods

2.1 Text collection

In this work, we used the TREC Genomics data in both 2006 and 2007 (Hersh et al. 2006, 2007), which consist of 28 topics and 36 topics (in the form of biological questions), respectively. In both years, a total of 162,048 full-text documents from Highwire Press[1] were used. Except for 1,800 instances, most of the documents were successfully mapped to their corresponding PubMed Identifiers (PMIDs). Hereafter, we refer to the remaining set of 160,248 PMIDs as the *TREC set* in our study. The 36 topics (Topic IDs 200 to 235) in 2007 and 28 (Topic IDs 160 to 187) in 2006 were collected from bench biologists and represent actual information needs in biomedical research. For demonstration purposes, we show three topics 207, 229 and 231 from TREC 2007 in the list below:

<207> What toxicities are associated with etidronate?
<229> What signs or symptoms are caused by human parvovirus infection?
<231> What tumor types are found in zebrafish?

For each topic, a set of relevant documents from the TREC set were produced by the relevance judgments based on pooled results from team submissions. They are assumed as the ground truth or gold standard data in our investigation and referred to as the *relevant set* in the remainder of this paper.

2.2 Query construction

For each TREC topic, a user query is required for retrieving relevant documents in PubMed. In order to produce unbiased (i.e. without human interference) user queries in a consistent manner, we chose to automatically select words from questions as queries on the

[1] http://highwire.stanford.edu/.

Table 1 Three potential user queries built automatically by selecting words from topic 229

Word combination	Recall	Precision	F-measure
human parvovirus	0.47	0.63	0.54
human parvovirus infection	0.39	0.85	0.53
parvovirus infection	0.39	0.69	0.49

assumption that real users would also intuitively create their queries based on the questions. Specifically, for each question, we first removed stop words (Wilbur and Sirotkin 1991) from the question and enumerated all possible word combinations, each of which was then expanded by ATM and subsequently searched in PubMed. Next, for each word combination that retrieved a non-empty set of documents, we compared those retrieved documents to the ones in the relevant set and computed the standard IR measures (Hersh 2003): recall, precision and F-measure (see also: Sect. 3.1). In the end, the query with the highest F-measure was selected.

Take the topic 229 for example, we first removed stop words *what*, *or*, *are*, *by* from the question. A total of 63 different user queries were then generated based on the remaining six words (*signs*, *symptoms*, *caused*, *human*, *parvovirus*, *infection*) and subsequently searched in PubMed to obtain a set of relevant documents. Three example queries are shown in Table 1, together with their corresponding IR measures after comparing the retrieved set with the relevant set. The query *human parvovirus* was finally chosen for further study because it yielded the highest F-measure. We processed all 64 topics in this way and were able to identify query terms for most of the topics except 9 instances (topics 164, 173, 177, 179, 180, 184, 185 in 2006 and topics 207 and 225 in 2007) where either no relevant documents were found in the gold standard or no query terms could be generated to represent meaningful user queries (i.e. using query expansion their F-measures are almost zero). Therefore, these nine topics were excluded from further analysis. The automatically generated queries for the remaining 55 TREC topics varied in length from a single word to a maximum of four words, with a mean of 2.5 words per query.

2.3 Preprocessing search field tags after translation

Once all of the 55 user queries were determined, their translated counterparts through PubMed's Automatic Term Mapping were obtained. For instance, *human parvovirus* was translated to: ((“humans”[TIAB] NOT Medline[SB]) OR “humans”[MeSH Terms] OR human[Text Word]) AND (“parvovirus”[MeSH Terms] OR parvovirus[Text Word]).[2] Translated queries may contain a variety of different search field tags. We filtered all but three tags: [MeSH Terms], [Text Words] and [All Fields] because they are (a) the most frequent tags seen after expansion (see also: Sect. 4.1); and (b) the major ones involving the use of MeSH. Therefore, they are the only important ones in our analysis for comparing results with respect to query expansion. For example, the function of the Subset search field ([SB]) is to restrict retrieval by topic, citation status and journal/citation subsets. Thus, the “human”[TIAB] NOT Medline[SB]” part of the translation for *human parvovirus* refers to searches for articles in PubMed but not yet indexed for MEDLINE (e.g. in-process citations which have been provided a PMID before it is indexed with MeSH). Since all of the citations in TREC 2006 and 2007 have already been indexed at the time of our

[2] Any changes to PubMed's ATM after March 2008 by the National Library of Medicine may result in slightly different translations as shown.

experiments, this specific search tag is not applicable to our data. Thus, it was discarded. As a result, the remaining part: "humans"[MeSH Terms] OR human[Text Word]) AND ("parvovirus"[MeSH Terms] OR parvovirus[Text Word]) served as the input to retrieve citations in our system.

All three search tags make use of the MeSH fields. However, the way they use the MeSH field varies. Based on the online help documents[3] and discussion with an expert medical librarian at the National Library of Medicine (Kathi Canese, personal communication), we treated the three search tags as follows in this work:

1. [Text Words]: Match against all words in MEDLINE title, abstract, and MeSH terms (no automatic MeSH explosion).
2. [MeSH Terms]: Match against the MeSH term and the more specific ones beneath it in the MeSH hierarchy (i.e. automated MeSH explosion).
3. [All Fields]: Same as [Text Words].

For the search tags [Text Words] and [MeSH Terms], they were handled identically in our analysis as they would be in PubMed. For [All Fields], we limited search fields to text words only because other bibliographic information of a MEDLINE citation (e.g. author name) was not used by our search system. A word or phrase will only be tagged with [All Fields] if no match can be found in existing translation tables. For instance, if a user query is *lysosomal*, it will be translated to lysosomal[All Fields] since no match can be found for it in any of the translation tables in PubMed.

As can be seen, PubMed would only retrieve documents matching the search keywords in the title or abstracts if the ATM was not used. After applying ATM, PubMed expands its search ability by examining the MeSH field of indexed documents, which could result in returning many relevant documents that would otherwise be missed. For each retrieved document in the following analyses, we were able to characterize whether it was retrieved because of the use of MeSH or because of word match in the title or abstract. In addition, for the ones that were retrieved because of MeSH, we further classified them into three different groups: matching a MeSH term, matching a more specific MeSH term due to automatic MeSH explosion, or matching words of a MeSH term (See also: Sect. 4.2). In the following experiments, query expansion refers to the inclusion of all three types of MeSH related retrievals.

3 Results

3.1 Evaluation by recall, precision and F-measure

Many different measures for evaluating the performance of IR systems have been proposed and used in evaluating the effectiveness of query expansion (Hersh 2003), two of which are selected in this study: *F-measure* and *mean rank precision*.

Retrieval performance was first assessed by recall (the proportion of relevant documents in the collection retrieved by the query) and precision (the proportion of retrieved documents relevant to the query). Then, F-measure was computed as the harmonic mean of recall and precision as follows:

[3] http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helppubmed.chapter.pubmedhelp.

Table 2 Results of average F-measures on the 2006 and 2007 TREC topics

TREC	Topics	Query expansion	Average F-measures
2006	21	Yes	0.406
		No	0.334
2007	34	Yes	0.264
		No	0.214

Table 3 Query-by-query F-measure analysis: with versus without query expansion (confidence level = 0.05)

TREC	Better	Worse	Equal	Statistical significance
2006	7	5	9	No
2007	20	8	6	Yes

$$F - measure = \frac{2 \times (precision \times recall)}{precision + recall} \quad (1)$$

As can be seen in Table 2, results using query expansion are better than those without the query expansion in terms of average F-measures in both TREC 2006 and 2007. Table 3 serves as a query-by-query F-measure analysis, describing the number of queries for which the query expansion technique achieved better, worse, or equal performance levels as done without the technique. The last column in Table 3 lists the results of the non-parametric statistical tests (Noreen 1989; Wilbur 1994) we performed through the pair-wise comparisons of all of the individual F-measures. Discrepancies between the results of 2006 and 2007 may be mostly attributed to the differences in topics and possibly the quality of relevance judgments. Our analysis shows that the relevant sets of TREC 2006 and 2007 consist of 1,448 and 2,478 documents, respectively. In contrast, about the same number of total relevant documents were retrieved (e.g., 661 in 2006 and 662 in 2007) when query expansion was used. As a result, performance (as shown in F-measures in Table 2) is considerably higher in 2006 than in 2007.

As the statistical test does not account for individual differences in magnitude, Fig. 1 and 2 show that for topics in TREC 2006 and 2007, the enhancements in F-measure are generally far greater than are the degradations. The largest performance improvement was observed to be over 7 fold. Without using query expansion, the F-measure for Topic 200 (see Table 4) in TREC 2007 was 0.047. With query expansion, the F-measure increased substantially to 0.398.

While we can see performance enhancement for most of the topics, query expansion resulted in degraded performance for some topics. The largest performance decrease happened to Topic 176 (see Table 4) where its F-measure dropped from 0.046 to 0.017

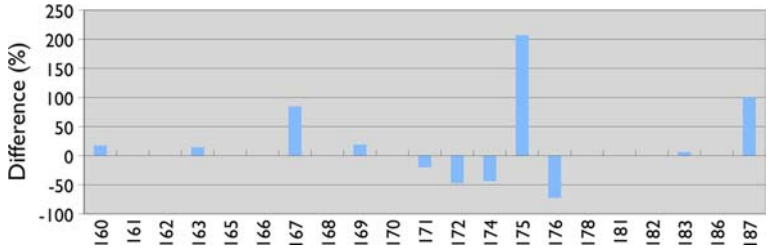

Fig. 1 TREC 2006: Query-by-query differences in F-measures

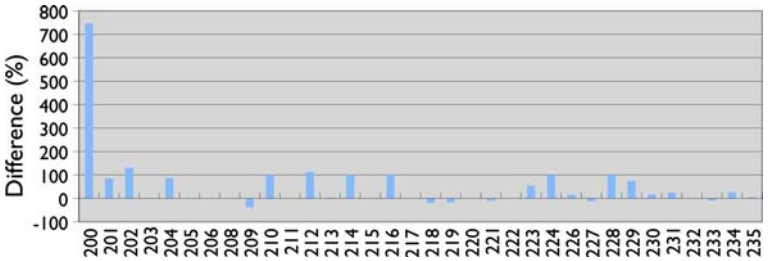

Fig. 2 TREC 2007: Query-by-query differences in F-measures

Table 4 The two topics (176 and 200) that display the largest performance improvement and degradation after applying the query expansion technique

Topic ID	Question	Constructed query
176	How does Sec61-mediated CFTR degradation contribute to cystic fibrosis?	CFTR degradation cystic fibrosis
200	What serum proteins change expression in association with high disease activity in lupus	Proteins lupus

(a 72.6% decrease) after applying the query expansion technique. Further analysis showed that the total number of retrieved documents increased over 10 fold (from 21 to 221) due to an expansion of the term: *degradation*. PubMed translated the text word *degradation* to "metabolism"[Subheading] OR ("metabolism"[TIAB] NOT Medline[SB]) OR "metabolism"[MeSH Terms] OR degradation[Text Word]. For this topic, this expansion was harmful to the performance.

3.2 Evaluation by mean rank precision

Our main concern in practice is to enhance retrieval performances among the top few returned documents which users are most likely to examine. To analyze this aspect, we introduce the second measurement for our experiments: Mean Rank Precision (MRP), which is the mean value of the rank precisions (precision at a given rank) computed for all queries. In this study, we chose the cut-off ranks to be 5, 10, and 20 as most of the retrievals occurred before 20 based on our own experience with PubMed users. Thus, the mean rank precision reveals more directly how a ranking strategy affects user retrieval effectiveness in realistic situations.

Note that PubMed returns citations in reverse chronological order by default, displaying more recent citations earlier. In accordance with this display order, we first sorted PMIDs numerically and then reversed the order. As a result, larger PMIDs would appear earlier in a ranked list.

Results in Table 5 show the mean average precisions at given ranks (5, 10, and 20) sorted by the reverse time order, suggesting the existing query expansion feature in PubMed does not always yield improved performance. In fact, the difference between results with and without query expansion is as small as a single relevant document for the TREC 2006 data. Additionally, results from TREC 2007 demonstrate that using such a technique could sometimes be harmful to system performance. This can happen because query expansion is generally considered as a recall-favoring technique (as opposed to precision-

Table 5 Results of the 2006 and 2007 TREC topics with and without query expansion as measured by mean rank precision

TREC	Query expansion	P@5	P@10	P@20
2006	Yes	0.556	0.540	0.540
	No	0.523	0.517	0.524
2007	Yes	0.385	0.411	0.413
	No	0.434	0.454	0.452

Retrieved documents are sorted in reverse chronological order

Table 6 Results of the 2006 and 2007 TREC topics with and without query expansion as measured by mean rank precision

TREC	Query expansion	P@5	P@10	P@20
2006	Yes	0.622	0.616	0.621
	No	0.613	0.592	0.584
2007	Yes	0.538	0.525	0.503
	No	0.566	0.548	0.519

Retrieved documents are sorted under the TF-IDF weighting scheme

favoring) that often results in a large number of additional retrievals. For instance, the number of retrieved documents increased from 1,174 to 3,163 when the query expansion technique was employed in TREC 2007. A much smaller increase in total number of retrievals (from 1,463 to 1,899) was seen in TREC 2006. This may account for the differences between 2006 and 2007, as shown in Table 5.

3.3 Results under different ranking strategy

We computed a new set of mean rank precisions when sorting retrieved documents under a different ranking strategy. Both previous research (Salton 1991; Salton and Buckley 1998) and in particular our own experience with PubMed (Lu et al. 1998) suggest that ranking by relevance can result in better retrieval performance. Thus, we computed TF-IDF scores for retrieved documents (Kim and Wilbur 2005; Lu et al. 2008) and then ranked them based on these scores. A document with a higher TF-IDF score is returned earlier in a list.

The same conclusions can be drawn with respect to the effect of using the query expansion technique based on the similar results shown in Tables 5 and 6. Superior results in mean rank precisions are seen in Table 6 compared with their counterparts in Table 5 due to a better ranking strategy. Together with experiments presented in Sect. 3.2, we confirmed an earlier expectation in (Hersh et al. 2000) where the authors believed that the difference in retrieval models (i.e. Boolean versus statistical-based) should not yield different results when evaluating query expansion techniques.

4 Discussion

4.1 PubMed's automatic term mapping process

For all the queries that we used in our work, no search tags were initially assigned. After automatic translation by PubMed's ATM, the majority of the queries were then associated

Table 7 Results of PubMed's automatic term mapping

Collection	Queries	[MeSH]	[Text]	[ALL]	[MeSH] OR [Text] OR [ALL]
TREC 2006	21	17	17	15	21
TREC 2007	34	28	28	23	34
Oct 17, 2005	2.4 M	1.5 M	1.4 M	1.4 M	2.0 M

The last column shows the number of queries assigned with at least one of the three search tags listed in the columns 3–5. Note that a single query can be assigned with multiple search tags (e.g. parvovirus[MeSH Terms] OR parvovirus[Text Words]). Thus for each row, the sum of the numbers in columns 3 to 5 is greater than the number in the second column (i.e. Queries). Abbreviations: [MeSH], [MeSH Terms]; [Text], [Text Words]; [ALL], [All Fields]

with either [MeSH Terms], [Text Words] or [All Fields] as shown in Table 7. Note that there could be multiple search tags associated with a single query after translation. Take the previous *parvovirus* example, its translation included both [MeSH Terms] and [Text Word] (i.e. "parvovirus"[MeSH Terms] OR parvovirus[Text Word]).

In order to show that the automatically generated queries are representative, we processed one day's worth of PubMed query data.[4] We discarded users that issued over 50 queries/24 hours as they are likely to represent programmatic searches. For the remaining 2,657,315 queries, we obtained their corresponding PubMed translations through the Entrez Programming Utilities (eutilities) (Geer and Sayers 2003). Results in the last row of Table 7 show that over 90% (2.4 M/2.65 M) of user queries are initially search tag free, over 80% (2.0 M/2.4 M) of which are expanded with at least one of the three tags studied in this work, suggesting that query expansion involving MeSH is a dominant phenomenon under real world circumstances. Other major search fields shown after translation include the ones in the following list. However, none of these search tags are related to MeSH terms. Therefore, they are beyond the scope of discussion in this paper.

1. TIAB NOT Medline[SB] (776,942): searching for terms in the title and abstract of articles in the databases that are not indexed for MEDLINE.
2. Author (380,628): searching for author names in the article.
3. Substance Name (89,510): searching for the name of a chemical in the article.

4.2 Different MeSH matching mechanisms

As we pointed out earlier in Sect. 2.3, the use of MeSH terms can be classified into two distinctive strategies:

1. If a query term is tagged with [MeSH Terms], the system will retrieve all documents indexed with this particular MeSH term as well as indexed with more specific MeSH terms beneath it in the MeSH hierarchy.
2. If a query term is tagged with either [Text Words] or [All Fields], our system will retrieve all documents indexed with one or more MeSH terms that include the query term. For instance, if a query term is *nervous system* and tagged with [Text Words] as 'nervous system'[Text Words], any documents indexed with the MeSH Term `Alcohol Abuse`, `Nervous System` will be retrieved in that 'nervous system' is part of this MeSH term.

[4] ftp://ftp.ncbi.nlm.nih.gov/toolbox/pubmed/query-logs/README.

Table 8 Performance (in F-measure) comparisons using different MeSH matching strategies

TREC	None	[MeSH Terms]	[Text Words] or [All Fields]	Both
2006	0.334	0.346	0.407	0.406
2007	0.214	0.260	0.236	0.264

Experiments were performed on both TREC 2006 and 2007 data

To characterize the individual contribution of each of the two classes to the overall improved performance, we performed two additional experiments where only one of the two matching mechanisms was exploited. Results are compared to each other in Table 8, and are also compared to the cases when none or both matching mechanisms were used. Results in Table 8 show that directly matching the MeSH term and its more specific terms is the major contribution for performance improvement in TREC 2007 while matching words in MeSH terms contributed significantly in TREC 2006. Therefore, we are unable to draw a conclusion on which one is more important for improving performance based on these experiments.

As a separate effort, we performed an additional experiment where we turned off the automatic MeSH explosion feature (i.e. more specific MeSH terms are no longer searchable) in the context of evaluating this specific feature. When compared to the results in the third column of Table 8, we see a performance drop from 0.346 to 0.334 in TREC 2006, and from 0.260 to 0.212 in TREC 2007, respectively. This leads us to conclude that the MeSH automatic explosion feature plays a critical role in the first type of MeSH usage.

4.3 Comparison with PubMed output

The main objective of this work is to discover useful features to improve PubMed search performance. In spite of the fact that our analysis was not based on output directly from PubMed, the retrieved document set by our custom search system was highly consistent with the actual PubMed output as shown in Fig. 3. The F-measures are almost identical for the systems. The slight discrepancies in performance come from two sources in our analysis: (a) we limited our searches with only three search field tags; and (b) for the search

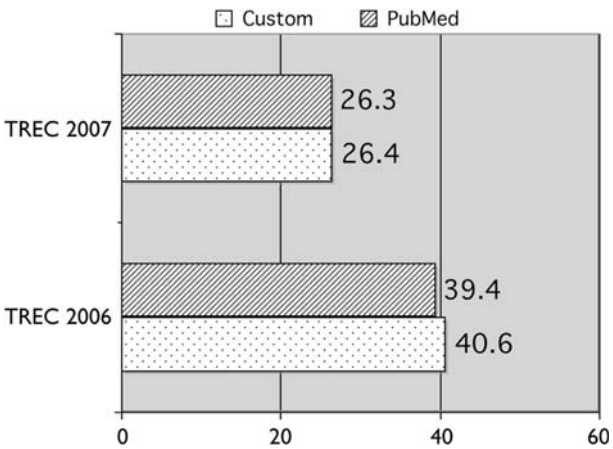

Fig. 3 The comparisons of F-measures based on outputs from PubMed vs. from our custom search system

tag [all fields], our search was limited to text words in title, abstract, and MeSH terms. Any other fields (e.g. author names) were not considered in our computation. The minor performance increase when using the custom-built system suggests (a) the three search tags that we selected to evaluate are critical for PubMed; and (b) that perhaps other search field tags could not further help to retrieve relevant documents for our query set.

5 Conclusions and future work

Based on the results of our large-scale analysis comprised of the 55 real biological questions and independently judged relevant documents, we conclude that query expansion using MeSH through PubMed's Automatic Term Mapping process can generally result in retrieving more relevant documents. However, this type of improvement may not prove useful for those users looking only at the top ranked returned documents (e.g. the first 20 returned documents).

The comparison results are useful in suggesting changes in PubMed (e.g. the display order). However, as we pointed out earlier, this work makes two assumptions. One assumption is that the automatically generated queries represent real user queries. The other assumption is that the documents in the relevant set are the ground truth and there are no more relevant documents in the TREC set. Our future research goal is to address both issues by involving human experts in evaluations.

Acknowledgments This research was funded by the Intramural Research Program of NIH, National Library of Medicine. The authors are grateful to the TREC organizers for their efforts in producing and making the text collection and relevance judgments publicly available.

References

Abdou, S., Ruck, P., & Savoy, J. (2005). Evaluation of stemming, query expansion and manual indexing approaches for the genomics track. In *Proceedings of the Fourteenth Text REtrieval Conference, Gaithersburg, MA*. Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Aronson, A. R. (1996). The effect of textual variation on concept based information retrieval. In *Proceedings of the AMIA Annual Fall Symposium* (pp. 373–377). Bethesda, MD: American Medical Informatics Association.

Aronson, A. R., & Rindflesch, T. C. (1997). Query expansion using the umls metathesaurus. In *Proceedings of the AMIA Annual Fall Symposium* (pp. 485–489). Bethesda, MD: American Medical Informatics Association.

Bacchin, M., & Melucci, M. (2005). Symbol-based query expansion experiments at TREC 2005 genomics track. In *Proceedings of the Fourteenth Text REtrieval Conference*. Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Carlin, B. G. (2004). Pubmed automatic term mapping. *Journal of the Medical Library Association, 92*(2), 168.

Gault, L. V., Shultz, M., & Davies, K. J. (2002). Variations in medical subject headings (mesh) mapping: From the natural language of patron terms to the controlled vocabulary of mapped lists. *Journal of the Medical Library Association, 90*(2), 173–180.

Geer, R. C., & Sayers, E. W. (2003). Entrez: Making use of its power. *Briefings in Bioinformatics, 4*(2), 179–184.

Guo, Y., Harkema, H., & Gaizauskas, R. (2004). Sheffield university and the TREC 2004 gemonics track: Query expansion using synonymous terms. In *Proceedings of the Thirteenth Text REtrieval Conference*. Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Hersh, W. (2003). *Information retrieval: A health and biomedical perspective* (2nd ed.). New York, NY: Springer-Verlag.

Hersh, W. & Bhupatiraju, R. T. (2003). TREC Genomics track overview. *The Twelfth Text Retrieval Conference, TREC 2003* (pp. 14–23). Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Hersh, W., Bhupatiraju, R. T., & Price, S. (2003). Phrases, boosting, and query expansion using external knowledge resources for genomic information retrieval. In *Proceedings of The Twelfth Text REtrieval Conference*. Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Hersh, W., Bhupatiraju, R. T., Ross, L., Johnson, P., Cohen, A., & Kraemer, D. (2006). TREC 2006 genomics track overview. In *Proceedings of the Thirteenth Text REtrieval Conference*. Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Hersh, W., Cohen, A. M., Roberts, P., & Rekapalli, H. K. (2006). TREC 2006 genomics track overview. In *Proceedings of the Fifteenth Text REtrieval Conference*. Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Hersh, W., Cohen, A. M., Roberts, P., & Rekapalli, H. K. (2007). TREC 2007 genomics track overview. In *Proceedings of the Sixteenth Text REtrieval Conference*. Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Hersh, W., Price, S., & Donohoe, L. (2000). Assessing thesaurus-based query expansion using the umls metathesaurus. In *Proceedings of AMIA Annual Symposium* (pp. 344–348). Bethesda, MD: American Medical Informatics Association.

Herskovic, J. R., Tanaka, L. Y., Hersh, W., & Bernstam, E. V. (2007). A day in the life of PubMed: Analysis of a typical day's query log. *Journal of the Medical Library Association, 14*(2), 212–220.

Kim, W., & Wilbur, W. J. (2005). A strategy for assigning new concepts in the MEDLINE database. In *Proceedings of AMIA Annual Symposium* (pp. 395–399). Bethesda, MD: American Medical Informatics Association.

Lindberg, D., Humphreys, B., & Mccray, A. (1993). The unified medical language system. *Methods of Information in Medicine, 32*(4), 281–291.

Lu, Z., Kim, W., & Wilbur, W. J. (2008). Evaluating relevance ranking strategies for medline retrieval. *Journal of the American Medical Informatics Association*, in press.

Mitra, M., Singhal, A., & Buckley, C. (1998). Improving automatic query expansion. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 206–214). New York, NY, USA: ACM.

Noreen, E. (1989). *Computer Intensive Methods for Testing Hypotheses: An Introduction*. Hoboken, NJ: John Wiley & Sons, Inc.

Robertson, S. E., & Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 232–241). New York, NY, USA. New York: Springer-Verlag.

Salton, G. (1991). Developments in automatic text retrieval. *Science, 253*(5023), 974.

Salton, G., & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing & Management, 24*, 513–523.

Salton, G., & Buckley, C. (1997). Improving retrieval performance by relevance feedback. In *Readings in Information Retrieval* (pp. 355–364). San Francisco, CA, USA: Morgan Kaufmann.

Shultz, M. (2006). Mapping of medical acronyms and initialisms to medical subject headings (mesh) across selected systems. *Journal of the Medical Library Association, 94*(4), 410–414.

Smith, A. M. (2004). An examination of pubmed's ability to disambiguate subject queries and journal title queries. *Journal of the Medical Library Association, 92*(1), 97–100.

Srinivasan, P. (1996). Query expansion and MEDLINE. *Information Processing & Management, 32*(4), 431–443.

Wilbur, W. (1994). Non-parametric significance tests of retrieval performance comparisons. *Journal of Information Science, 20*(4), 270–284.

Wilbur, W. J., & Sirotkin, K. (1991). The automatic identification of stop words. *Journal of Information Science, 18*(1992), 45–55.